AD_____

Award Number: DAMD17-02-1-0054

TITLE: Protein BioChip Mass Spectrometry: A Novel Protein Profiling Approach for Diagnosis and Prognosis of Prostate Cancer

PRINCIPAL INVESTIGATOR: Bao-Ling Adam, Ph.D.

CONTRACTING ORGANIZATION: Medical College of Georgia
Research Institute, Incorporated
Augusta, GA 30912-4810

REPORT DATE: January 2005

TYPE OF REPORT: Annual

PREPARED FOR: U.S. Army Medical Research and Materiel Command
Fort Detrick, Maryland 21702-5012

REPORT DOCUMENTATION PAGE

Form Approved
OMB No. 074-0188

1. AGENCY USE ONLY	2. REPORT DATE	3. REPORT TYPE AND DATES COVERED
	January 2005	Annual(1 Jan 2004 - 31 Dec 2004)

4. TITLE AND SUBTITLE
Protein BioChip Mass Spectrometry: A Novel Protein Profiling Approach for Diagnosis and Prognosis of Prostate Cancer

5. FUNDING NUMBERS
DAMD17-02-1-0054

6. AUTHOR(S)
Bao-Ling Adam, Ph.D.

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)
Medical College of Georgia
Research Institute, Incorporated
Augusta, GA 30912-4810

E-Mail: badam@mcg.edu

8. PERFORMING ORGANIZATION REPORT NUMBER

9. SPONSORING / MONITORING AGENCY NAME(S) AND ADDRESS(ES)
U.S. Army Medical Research and Materiel Command
Fort Detrick, Maryland 21702-5012

10. SPONSORING / MONITORING AGENCY REPORT NUMBER

11. SUPPLEMENTARY NOTES
Original contains color plates: ALL DTIC reproductions will be in black and white

12a. DISTRIBUTION / AVAILABILITY STATEMENT
Approved for Public Release; Distribution Unlimited

12b. DISTRIBUTION CODE

13. ABSTRACT (Maximum 200 Words)

The present study was designed for the early detection of prostate cancer and monitoring the outcome of prostatectomy treatment. All the prostate cancer patients had undergone radical prostatectomy with no positive nodes, no extra capsular extension, no positive seminal vesicle involvement at time of surgery, and no positive CT scans post surgery. This allowed us to build a pattern matching algorithm to detect early stage cancer. Currently, we are developing an in-house algorithm to preprocess the raw spectra data which includes three steps: (1). Background subtraction and noise reduction (2) Local alignment of spectra, (3) Peak detection. Developing a classification algorithm is in progress. This will be completed in Fall 2005. The completion of the data analysis will be expected in the end of 2005.

14. SUBJECT TERMS
Prostate Cancer, protein profiling, biomarker discovery

15. NUMBER OF PAGES
16

16. PRICE CODE

17. SECURITY CLASSIFICATION OF REPORT	18. SECURITY CLASSIFICATION OF THIS PAGE	19. SECURITY CLASSIFICATION OF ABSTRACT	20. LIMITATION OF ABSTRACT
Unclassified	Unclassified	Unclassified	Unlimited

Table of Contents

INTRODUCTION

This report is a summary of the study conducted in Dr. Adam's laboratory at The Medical College of Georgia from January 2004 to December 2004.

Since Dr. Yutaka Yasui relocated and withdrew from this study, Dr. Ovidiu Lipan took this position to assist in data analysis. Dr. Ovidiu has developed a new data analysis strategy which is reported here.

The specific aims of this study are to 1) evaluate the clinical utility of SELDI fingerprint protein profiling as a diagnostic test for prostate cancer, 2) evaluate the application of SELDI fingerprint protein profiling for prognosis of prostate cancer.

The outline below is to identify the portion of the project that has been completed. The Blue ink indicates the proposed task; the black ink indicates the progress.

Task 1. Evaluate the clinical utility of SELDI fingerprint protein profiling as a diagnostic test for prostate cancer (months 1-24).

a) We will process 600 serum samples from 200 patients diagnosed with prostate cancer, 200 patients with benign prostate disease and 200 age-matched normal men, using the SELDI ProteinChip® system (months 1-6).

We have processed 197 patients diagnosed with prostate cancer (PCA), 93 patients with benign prostate hyperplasia (BPH) and 96 age-matched normal men, using the SELDI ProteinChip® system. This is the first study.

b) The pre-processed SELDI data obtained from the processing of the above prostate samples will be used to develop and train a Wavelet Transform/Information learning algorithm (months 2-12).

The Time-of-Flight data obtained from the processing of the above prostate samples were successfully used to develop and train a Wavelet Transform/Information learning algorithm as well as other learning algorithms.

c) Diagnostic criteria will be established and applied to the learning algorithm. Specially, we will establish the diagnostic cutoff point, the specificity/sensitivity of the learning algorithm for detecting prostate cancer, the receiver operating characteristics (ROC) curve to evaluate the efficiency of the diagnostic test, and the reproducibility of the SELDI protein profiling assay (months 7-12).

This step was successfully completed and the results were presented in the progress report of 2002.

d) A validation study of the diagnostic algorithm will be conducted on 1200 serum samples, 400 in each category to assess the specificity/sensitivity for discriminating prostate cancer from benign prostate disease and normal healthy men. The SELDI data will be compared to DRE, serum PSA level and pathological stage (months 10-24).

A validation study (second study) was started with 181 PCA, 143 BPH, 199 age-matched normal (age> 50) and 123 normal young (age < 50). These samples (total 1938 serum samples) have been processed for SELDI reading. SELDI data has been collected. The data has been compiled and the data analysis is in progress. **The part of data analysis results will be presented in this report.**

e) The diagnostic algorithm will be evaluated for the possible interference from other diseases by testing cancer of different types, benign urological conditions, and diseases such as hypertension, and diabetes. We will analyze approximately 50 sera in each of these categories, the total 400 serum samples will be analyzed (months 10-24).

We are recruiting samples from other cancer types and other diseases. We have completed the recruiting of 50 breast cancer, 50 leukemia, 50 head and neck cancer, 50 bladder cancer, and 50 liver cancer samples. **These samples are ready for the assay which will be conducted in 2005.**

Task 2. Application of the SELDI fingerprint protein profiling for the prognosis of prostate cancer (months 20-36).

a) The same diagnostic algorithm will be used to evaluate the protein profiles obtained from pre- and post-treatment serum samples. It will allow us to develop the prognostic indicator to decide the type of treatment, assess adequacy of a given therapy, and correlate closely with evidence of disease progression or recurrence (months 20-36).

We have conducted a study with 150 cancer patients with pre- and post-prostatectomy serum samples. The data has been collected and the data analysis is in progress. After the initial data analysis, we realized that the SELDI instrument was not calibrated correctly, therefore, we are in the process of re-running these pre-and post samples.

b) The correlative value of tumor profiles in relation to different therapies will be evaluated. We will analyze 600 samples from prostatectomy patients and 100 samples each for radiation and hormone ablation patients (months 20-36).

We have started recruiting samples. We have collected 80 samples so far. It is far from what our plans; therefore, I have established collaboration with Dr. John Semmes at Eastern Virginia Medical School. He will provide me some of the samples I need from his biorepository. This should move this aim forward in the end of this year.

d) Evaluate the efficiency of our protein profiling as a prognostic indicator to existing prognostic indicators such as PSA, pathologic stage and grade of prostate cancer (months 28-36).

It will be the focus of year 02 at MCG.

e) Classification of prostate cancer into clinically defined groups using signature protein profiles (months 28-36).

It will be the focus of year 02 at MCG.

f) Identification of "pre-metastatic" prostate cancer via signature protein profiles (months 28-36).

It will be the focus of year 02 at MCG.

BODY

Human Assurance Committee

We have received the IRB approval through the MCG Human Assurance Committee and have started the sample collection.

Sample Recruitment and Demographics:

We started recruiting samples for this second study to validate what we observed form the first study. In order to design a race, age balanced, well designed study, a total of 1300 non-cancer normal donor, 195 prostate cancer (PCA) patients, and 240 benign prostate hyperplasia (BPH) patients were obtained from our serum bank. Based on the race and age, the final sample size for normal, BPH, PCA were 321, 142, and 181, respectively.

Eligibility criteria for patients to this study are shown in **Table 1** of the appendix. In order to select early stage cancer specimens for this study, only pre-treatment serum from the clinically localized prostate cancer patients were selected. To fit into clinical localized prostate cancer criteria, the patients have to undergo radical prostectomy. There is no positive nodes found at time of surgery and no positive CET scans found post surgery. There is no extraccapsular extension and no positive seminal vesicle involvement found in the patients. It is very difficult to define BPH group. Therefore, we selected patients who were biopsy proven BPH. They were also pathologically confirmed that there was no cancer found. The normal population includes young group (younger than 50 years of age) and age matched group (older than 50 years of age). For normal healthy donors, we requested donors with a PSA value less than 4, normal DRE and no known cancer or other urinary disease. The PSA was not used as a criterion in BPH and PCA groups, but only in both normal young and age-matched group. The PSA distribution in study population is shown in **Figure 1, Appendix**. The range and mean PSA values for the young normal group was 0 to 4 ng/ml (0.95 ± 0.62 ng/ml); 0 to 3.89 ng/ml (1.33 ± 0.86 ng/ml) for age-matched normal group; 0.2 to 30.9 ng/ml (6.33 ± 5.66 ng/ml) for the BPH group; and 0 to 196 ng/ml (7.97 ± 16.00 ng/ml) for PCA group.

Table 1. The criteria of sample selection for this study were decided to assure a well designed study.

Criteria for Sample Selection

Clinically localized prostate cancer: N=186
- Radical prostectomy patients
- Pre-treatment samples
- No positive nodes at time of surgery
- No positive CT scans post surgery
- No extracapsular extension
- No positive seminal vesicle

Benign prostate hyperplasia (BPH): N=142
- Biopsy proven BPH which was pathologically confirmed there is no cancer either by biopsy or TURP

Age-matched Normal: N=219
- PSA is less than 4
- Normal DRE
- No known cancer or other urinary disease
- Donors are selected also based on the age, race distribution of both BPH and cancer group

Young Normal: N=102
- Age younger than 50
- PSA is less than 4
- Normal DRE
- No known cancer or other urinary disease
- Donors are selected also based on the race distribution of both BPH and cancer group

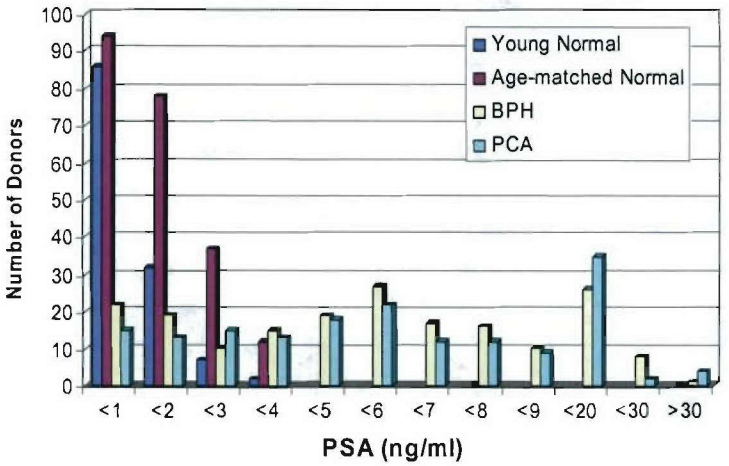

Figure 1. PSA distribution among four different groups in this study (Young Normal, Age-matced Normal, BPH, and PCA)

The age-matched normal group consisted of 39 African Americans and 182 Caucasians with an age range from 50 to 74 (medium 64). There were 80 African Americans and 47 Caucasians in the normal young group ranging in age from 30 to 49 (medium of 45). The BPH group had 80 African American and 171 Caucasians with an age range from 51 to 79 (medium of 67). The PCA group consisted of 34 African Americans and 136 Caucasians ranging in age from 50 to 74 (medium of 62). (Figures 2 and 3, Appendix). Overall the experimental design is considered well balanced in age and race as much as possible.

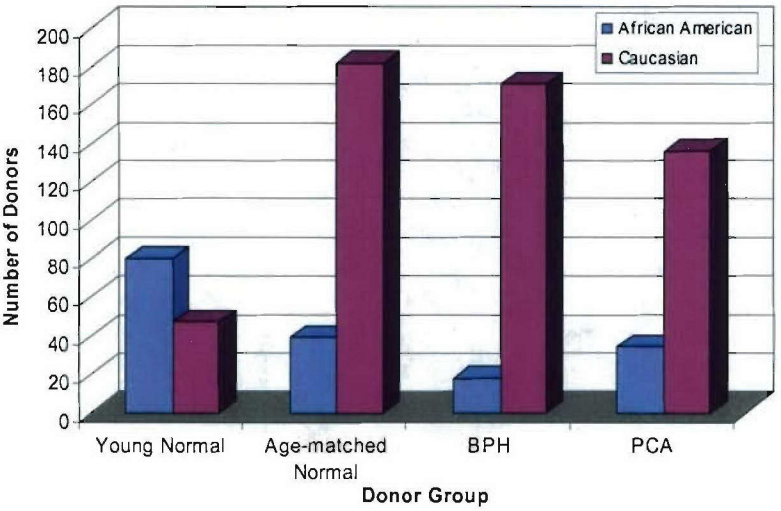

Figure 2. Race distribution of each donor group.

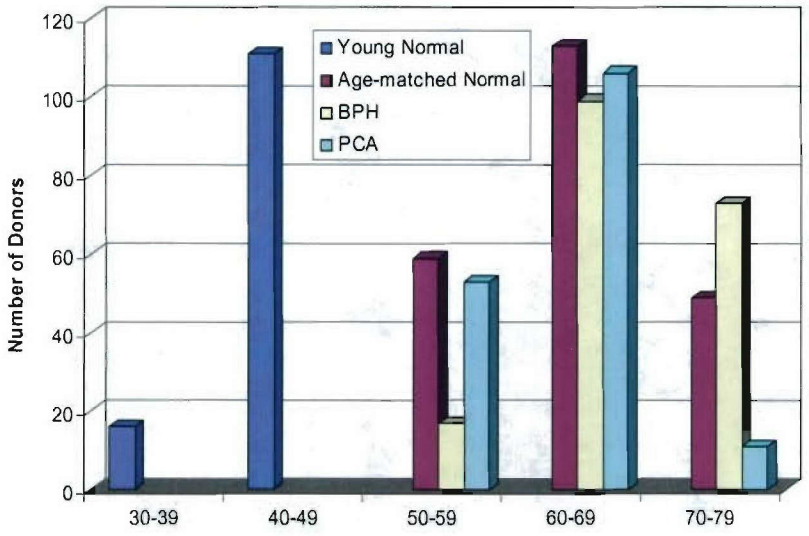

Figure 3. Age distribuation of each donor group.

Sample processing:

IMAC-3 chips (Ciphergen Biosystems, Inc, Fremont ,CA) were coated with 20 ul of 100 mM CuSO4 on each array, placed on a TOMY Micro Tube Mixer (MT-360, Tomy Seiko Co., Ltd), and agitated for 5 minutes. The chips were rinsed with deionized (DI) water 10 times, 20 ul of 100 mM sodium acetate were added to each array, and shaken for 5 minutes to remove the unbound copper. The chips were rinsed again with DI water for 10 times and put into a bioprocessor (Ciphergen Biosystems, Inc.), which holds 12 chips, and allows for applying larger volumes of serum to each chip array. The bioprocessor was washed and shaken on a platform shaker at a speed of 250 rpm for 5 minutes with 200 ul PBS in each well. This was repeated twice more and each time the PBS buffer was discarded by inverting the bioprocessor on a paper towel. Serum samples for SELDI analysis were prepared by vortexing 20 ul of serum with 30 ul of 8M Urea/ 1% CHAPS in PBS a 1.5 ml microfuge tube at 4°C for 10 minutes. 100 ul of 1M urea with 0.125% CHAPS was added into the serum/Urea mixture and briefly vortexed. PBS was added to make a 1:5 dilution which was placed on ice until applied to a protein chip array. 50 ul of the diluted serum-Urea-mixture was applied to each well, the bioprocessor sealed, and shaken on a platform shaker at a speed of 250 rpm for 30 minutes. The serum-Urea mixture was discarded and the PBS washing step was repeated 3 times. The chips were removed from the bioprocessor, washed with DI water 10 times, and air-dried. The chips were stored in the dark at room temperature until subjected to SELDI mass analysis. Prior to SELDI analysis, 0.5 ul of a saturated solution of the energy absorbing molecule (EAM) sinapinic acid (Ciphergen Biosystems, Inc.) in 50% (v/v) acetonitrile, 0.5% trifluoroacetic acid was applied onto each chip array twice, letting the array surface air dried between each sinapinic acid application. Chips were placed in the Protein Biological System II (PBS II) mass spectrometry reader (Ciphergen Biosystems, Inc.), and time of flight mass spectra were generated by averaging 192 laser shots collected in the positive mode at laser intensity 220, detector sensitivity 7, and focus lag time of 900 ns. Mass accuracy was calibrated externally using the All-in-1 peptide MW standard (Ciphergen).

Data Analysis:

The data analysis is currently in progress. The preliminary results were obtained using an in house SELDI program. The data analysis was divided into four different permutations (age-matched normal vs. BPH; age-matched normal vs. PCA; age-matched normal vs. normal young; BPH vs. PCA). These results were reported in 2004 report.

We have finished more data analysis based on Dr. Yasui's method. This data analysis was done on three different comparisons: age-matched control vs. PCA, age-matched control vs. BPH and PCA, BPH vs. PCA.

Training Data	Sensitivity	Specificity
Age-matched normal vs. PCA	**77%**	**89%**
Age-matched normal vs. BPH and PCA	**87%**	**84%**
BPH vs. PCA	**70%**	**62%**

From these preliminary data, we see the potential to distinguish PCA from age-matched normal patients. Although the sensitivity and specificity could be better, we are confident further fine tuning of the data analysis and better data processing will improve the performance. These are the results from Dr. Yasui's data analysis.

We are continuing data analysis and algorithm development with Dr. Ovidiu Lipan on MCG campus. The descriptions below are the methods developed by Dr. Lipan. We first developed a scheme to preprocess the raw spectra data which includes three different steps: (1). Background subtraction and noise reduction (2) Local alignment of spectra, (3) Peak detection.

Raw spectra preprocessing

1. Background subtraction and noise reduction

The method to evaluate the baseline is based on Sensitive Nonlinear Iterative Peak clipping (SNIP) algorithm, [1]. We calculate the baseline of the spectrum $S(m)$, which mathematically is represented as a vector. From $S(m)$ we calculate step by step a series of vectors $S_1(m)$, $S_2(m)$ up to $S_K(m)$, where K is a parameter fixed by the user. The new value in the p^{th} iteration step

is obtained by comparison of the average of the values $S_{p-1}(m-p)$, $S_{p-1}(m+p)$ with the value $S_{p-1}(m)$. We accept the minimum of these values, i.e.,

$$S_p(m) = min\left(S_{p-1}(m), \frac{S_{p-1}(m-p) + S_{p-1}(m+p)}{2} \right)$$

By taking the minimum, we make certain that the background stays beneath the spectrum, **Figure 4B.**

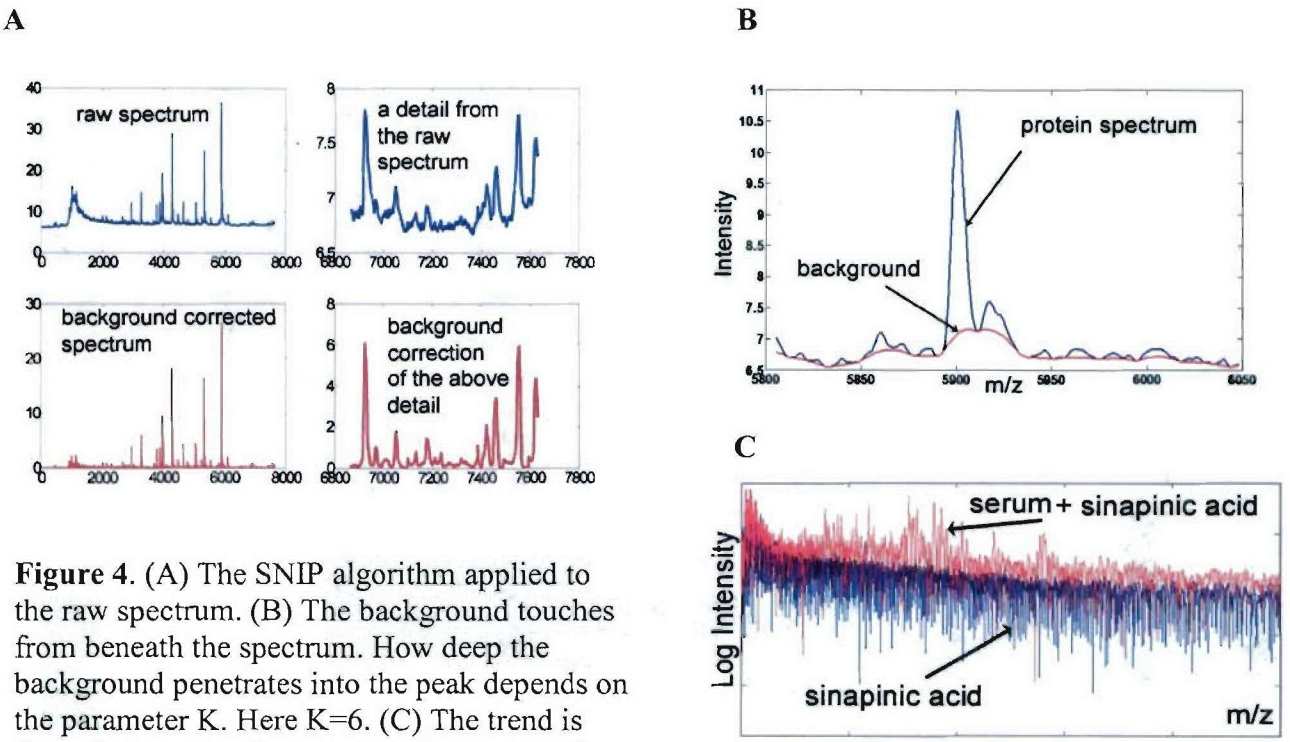

Figure 4. (A) The SNIP algorithm applied to the raw spectrum. (B) The background touches from beneath the spectrum. How deep the background penetrates into the peak depends on the parameter K. Here K=6. (C) The trend is linear in logarithmic scale. The trend effect is visible with or without serum.

The SNIP algorithm is designed to follow the local behavior of the spectrum so that the baseline will chase the local trend. On top of the local trend, we noticed the existence of a global trend in each spectrum, **Figure 4C**. This global trend has a universal property: in logarithmic scale it decays linearly. To confirm that this effect is not biological, we plotted two spectra in Figure 10C. One spectrum contains the serum and the Sinapinic Acid (red), and the other spectrum contains only the Sinapinic Acid (blue). Note that the Sinapinic Acid is used as control, because this acid is loaded first on the chip and then the neat serum is added. In both cases the global trend is present. We corrected this trend by fitting a straight line through the spectrum in logarithmic scale. The fitting procedure avoids using the peaks and the noise, by employing the spectral intensities that lie between 1st and 3rd quantile of the spectral values. After the background and the global trend are corrected, the next step is to eliminate the peaks that are noise fluctuations.

The procedure we used is a version of the SNIP algorithm. We already know that the background is a curve that flows beneath the spectra and touches it at the local minima. A filtering method should eliminate the small peaks that are highly variable and keep the strong peaks untouched. If instead of the spectrum $S(m)$ we use as an input $1/S(m)$ in the SNIP algorithm we obtain a curve that follow the spectrum from above and this curve will avoid small valleys and peaks that constitute the noise. As the number of iteration increase, more and more of the small peaks are eliminated, **Figure 5**.

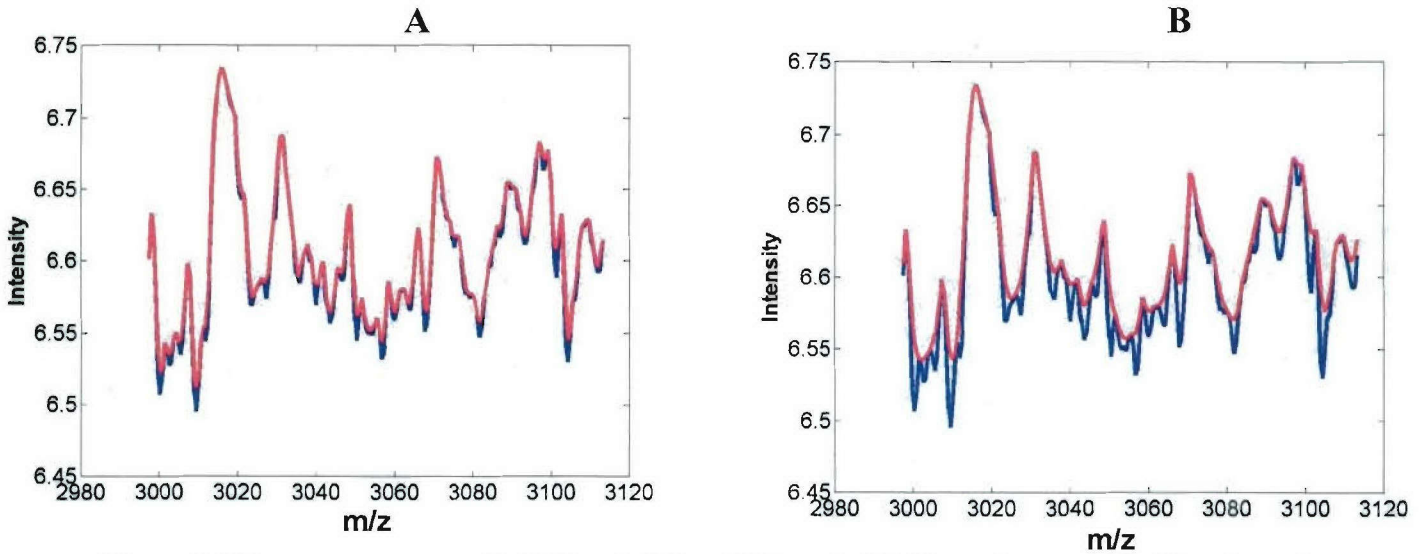

Figure 5. Filter process was applied 1 time in (A) and 6 times in (B). The red curve is the filtered spectrum.

2. Local alignment of spectra

Once each individual spectrum has passed the first part of the data processing, the whole group of the spectra must be aligned. In other words, the spectra must be stacked one on top of the other to check if the same protein corresponds to the same mass/charge. To insure a proper alignment of spectra, the SELDI instrument is calibrated with the help of a set of proteins with known mass/charge values. This procedure is efficient for a global alignment of the spectra. However, as we see from the **Figure 6A**, locally, the spectra are still out of phase. To correct this effect, regions of spectra must be shifted against each other. Two challenges must be overcome in determining the right shifts: 1) the shifts are mass/charge dependant and 2) we do not have a reference spectrum to align the other spectra against it. The mass/charge dependence of the shifts is solved by dividing the entire mass/charge region in a number of M pieces. If the region is divided in too many pieces, M = 256 for example, then the alignment will be influenced by noise. If M=4, then the regions are too large and we loose the mass/charge dependence of the shifts. We used M = 16 for mass/charge values between 3kDA and 25kDA. For each of the 16 pieces we must compute the shift $z(p,S_j)$ where $p=1...16$. Here the index $j=1...N$ represents the spectra to be aligned. Because we do not have a reference spectrum, we use all pairs of spectra as the starting point of our procedure. For a pair of spectra (S_j,S_k) and a mass/charge region p we determine $\tau_{jk}(p)$ from the condition that the Euclidian distance between the

spectra $S_j(m)$ and $S_k(m+\tau_{jk}(p))$ is minimum (here $m \in p$). After all $\tau_{jk}(p)$ for $j>k$ are determined the shifts $z(p,S_j)$ are estimated using the hypothesis that each τ is a difference of two shifts:

$$\tau_{jk}(p) = z(p,S_j) - z(p,S_k)$$

For a set of N spectra there are $N(N-1)/2$ numbers τ_{jk}. To estimate N shifts $z(p,S_j)$, $j=1...N$, from $N(N-1)/2$ values for τ, we use a linear regression procedure. This procedure proved to be very efficient for local alignment, (**Figure 6B**).

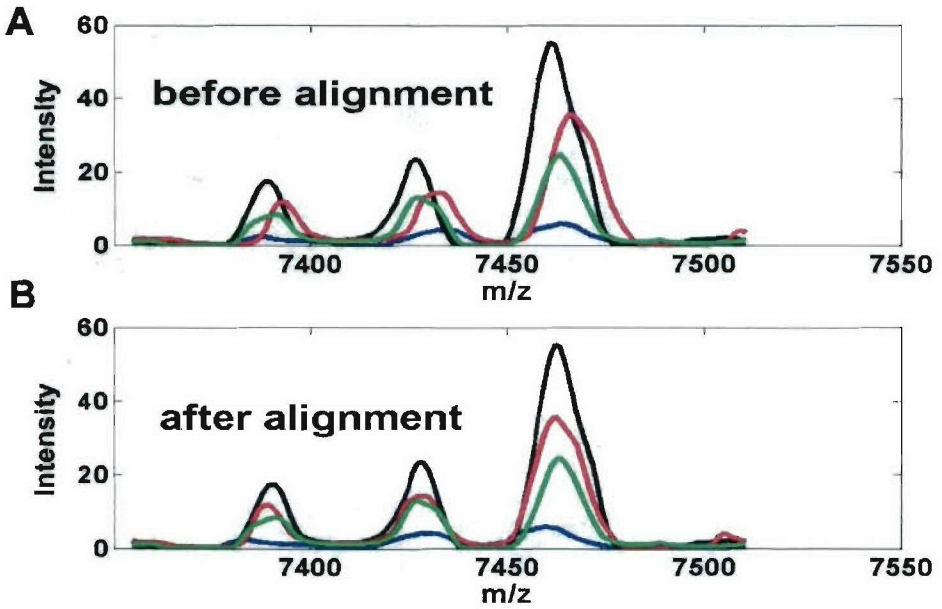

Figure 6. The set of spectra before (A) and after (B) the alignment procedure.

3. Peak detection

The apparently simple task, to the eye, of selecting narrow peaks that rise significantly above a baseline is not straightforwardly transformed into an algorithm. The problem is to distinguishing true peaks from statistical fluctuations. The presence of noise in the spectra results in committing two types of errors: a type I error in which a strong fluctuation is detected as a real protein peak and a type II error in which small real protein peaks are not detected. The peak identification method should minimize the probability of committing a type I error and, at the same time, should maximize the probability of detecting small peaks. The peak detecting procedure used in spectra similar to proteomic ones (gamma-ray spectra, nuclear spectra) are based on a linear transformation (convolution) of the experimental data:

$$\overline{S}(i) = \sum_{k=i-m}^{i+m} c(i-k)S(k),$$

with $c(k)$ real constants obeying the constrains

$$\sum_{k=-m}^{m} c(m) = 0,\ c(k) = c(-k).$$

The coefficients $c(k)$ actually form a filter. The number of these coefficients $2m+1$ must be chosen by the user. We used $m = 64$, which is large enough to detect strong peaks and small enough to detect peaks that are in the proximity of a high peak. Diverse filter coefficients $c(k)$ will give different outcomes for the probabilities of type I and type II errors. If we assume that the shape of the peak is Gaussian, then the best filter is the so called matched filter given by

$$c(k) = g(k) - \frac{1}{2m+1}\sum_{j=-m}^{m} g(j),$$

$$g(j) = exp\left(-\frac{j^2}{2\sigma^2}\right),$$

with σ being the width of the peak to be detected. The detection is based then on the ratio

$$f(i) = \frac{\sum_{k=i-m}^{i+m} c(i-k)S(k)}{\sqrt{\sum_{k=i-m}^{i+m} c(i-k)^2 S(k)}}$$

which is computed out of the data $S(k)$ and the filter coefficients, **Figure 7.** The probability of detecting false peaks can be expressed only in terms of the ratio f, [2]. For $f = 3$ the probability of detecting false peaks is 0.00025. We will consider that a peak was located at those local maxima of the sequence $f(i)$ that are above the critical value 3 (the index i in $f(i)$ represents the mass/charge, so $i = m/z$).

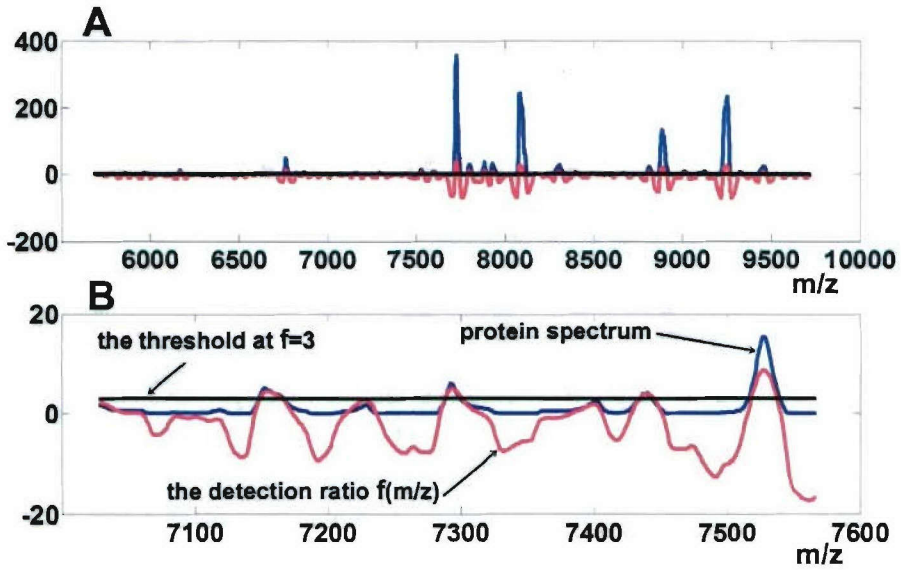

Figure 7. The blue curve is the protein spectrum; the red curve is the selection ratio f(m/z); the black line is the f = 3 selection value. The spectrum peaks are selected when the red curve goes

above the f = 3 line. The mass is measured in Da. (A) presents a large portion of the spectrum, whereas (B) concentrates on a small mass/charge zone.

The probability that a small peak is not detected (type II error) depends on the ratio f, but also on the height of the peak, [2]. Consider a Gaussian peak at position i_0 with a height S on top of the constant background R (what is left after the baseline subtraction),

$$S\,exp\left(-\frac{(i-i_0)^2}{2\sigma^2}\right)+R$$

The probability of not detecting small real peaks depends on the parameter $\alpha = S/\sqrt{R}$ that measure the relative intensity of the peak with respect to the background. The probability of missing very small real peaks ($\alpha = 1$) is 0.9 if we choose $f = 3$ as above. Evidently, if f is lowered to $f = 0.1$ the probability of detecting very small peaks rises to 0.7. However, the probability of detecting false peaks increases to 0.1. These facts lead us to avoid searching for low intensity peaks. Fortunately, the probability of detecting peaks with $\alpha = 2,3$ is between 0.8 and 0.9 for $f = 3$. Usually we select about 200 peaks and so we miss about 30 small peaks without contaminating the selected peaks. (The number of false detected peaks is $0.00025 \cdot 200 = 0.05$.)

We finished about 80% of this algorithm development, we are in the stage of organize the codes of this algorithm to improve the performance and write up the user manual. This will be completed in next three months. In the mean time, we are developing the classification scheme to discriminate the normal form PCA and BPH.

KEY RESEARCH ACCOMPLISHMENTS

1. Collected specimens for this study.
2. Established the collaboration with EVMS to facilitate this study.
3. Developed the mathematic algorithms for proteomic data analysis

REPORTABLE OUTCOMES

1. Two manuscripts were published.

Wagner M, Naik DN, Pothen A, Kasukurti S, Devineni RR, **Adam BL**, Semmes OJ, Wright GL Jr. Computational protein biomarker prediction: a case study for prostate cancer. BMC Bioinformatics. 2004 Mar 11;5(1):26.

Yasui Y, Pepe M, Hsu L, **Adam BL**, Feng Z.Partially Supervised Learning Using an EM-Boosting Algorithm. Biometrics. 2004 Mar;60(1):199-206.

CONCLUSIONS

The SELDI technology has successfully established in Dr. Adam's lab at the Medical College of Georgia. The specimen recruitment will take some time to achieve the goal for this study. Dr. Adam has established the collaboration with Dr. John Semmes to facilitate this study. Therefore, this study should be in the right time course as planned. With the progress of Dr. Lipan's algorithm development, the data analysis will be on its way in the Fall of 2005.

REFERENCES

1. Ryan CG, Clayton E, Griffin WL, et al. SNIP, A statistics-sensitive background treatment for the quantitative-analysis of pixe spectra, Geoscience Applications. *NUCL. INSTRUM. METHODS* B 34 (3): 396-402 Sep 1988

2. Hnatowicz V. Identification of weak lines in gamma-ray spectra. *NUCL. INSTRUM. METHODS.* 133 (1): 137-141 1976